Res. Found., Washington), Vol. 5, Suppl. 3, pp. 345-358 (1978)). The mutation probability matrix PAM 128, generated from the PAM1 matrix as described by Dayhoff, was used.

In the Gribskov method, the value of the profile for amino acid *a* at position *p* is given by

$$M(p,a) = \sum_{b=1}^{20} W(p,b) \times Y(a,b)$$

where Y(*a,b*) is the probability obtained from Dayhoff's mutation probability matrix for the substitution of *a* for *b*, and W(*p,b*) is a weight for amino acid b at position p.

The frequency of an amino acid in the alignment at a particular position was used for its weight:

$$W(b,p) = n(b,p) / N_r,$$

where *n*(*b,p*) is the number of times *b* appears at position *p*, and $N_r$ is the total number of amino acid counts at that position.

The probability matrix, GG36 residues against all 20 substitution residues, was normalized to the largest fraction in each row. *See* Table 2.

The constraint vector was designed such that mutagenesis would focus on positions which are close to the active site of the enzyme. The calculation was based on two crystal structures which have peptides bound to different regions of the active site: a structure of FN2 (a subtilisin mutant from *B. lentus*, which is identical to GG36 except for the following substitutions; K27R, V104Y, N123S, and T174A) which contained the peptide Ala-Ala-Pro-Phe bound to the $S_4$ to $S_1$ subsites; and a structure of subtilisin BPN' (from *B. amyloliquefaciens*) which had the inhibitor Suc-Ala-Phe-Ala bound to the $S'_1$ to $S'_3$ subsites. Both structures were aligned using the program "insight II" (MSI, San Diego, CA). Subsequently, the coordinates of the inhibitor Suc-Ala-Phe-Pro-Ala were moved into the structure of FN2. The combined coordinates were imported into Excel (Microsoft, Redmond, WA). For each residue of the enzyme the distance between the beta carbon atom and the closest beta carbon atom of the two bound peptides was calculated. Where glycine residues, which do not have a beta carbon, occurred, the distance between the alpha carbon of the glycine residue and the beta carbon of the bound peptide was calculated instead.

For each backbone residue, a selection value was calculated using the constraint vector as described below. This value was used to select residues from the sequence profile for

inclusion in the substitution table. Profile values greater than or equal to the selection value were added to the substitution list for that position. The lower the value, the increased chance that a substitute residue was selected at that position.

A linear constraint vector of the formula y=mx+b was used to generate the combinatorial selection scheme, where x=Cβmin. The m and b terms were chosen to provide ~100 substitutions from residues between 1 and 10 Å from the active site as described, yielding m=0.15500 and b=-0.40000. Any y values >1 (which result from a distance of >10 Å) were ignored. Entries in the profile shown in Table 1 which exceeded the y value determined for that position by applying the constraint vector (and <1) were selected for inclusion in the combinatorial library. Application of the constraint vector to the probability matrix in this manner produced the substitution table shown in Table 3, containing 105 suggested substitutions.

Visual inspection of the enzyme structure determined that most residues which are close to the bound ligand were included in the mutagenesis scheme. It was decided to avoid mutation of positions H62 and S215 as proposed by the algorithm because these two residues are part of the catalytic triad of subtilisin. Furthermore, V66C was eliminated from the mutagenesis scheme because an unpaired Cys residue is unlikely to lead to a functional GG36. These alterations represent contribution of a knowledge-based constraint to the results produced by applying the constraint vector to the probability matrix. As the consensus sequence derived from alignment of the large family was quite different from that of savinase, the most prevalent residue at several positions in the profile was not the residue in the savinase backbone. Additionally, in some cases the wild type residue was suggested to be substituted with itself. In cases where only a single substitution of a residue was suggested, the technique used to form the library could be doped with the wild type residue to prevent inclusion of a possibly debilitating residue in all members of the library.

Example 2. Alteration of β-lactamase Specificity Using a Scoring Profile

This example demonstrates the application of a distance-based constraint vector to a position-specific scoring matrix generated using a multiple sequence alignment of seven members of the ampC family of proteins and a PAM32 substitution matrix.

To create the IRL produced in this example, 7 beta lactamase ampC protein sequences (those from *A. sobria*, *E. coli*, *O. anthropi*, *P. aeroginosa*, *S. enteriditis* and *Y. enterolitica* ) were aligned using the default parameters of the program AlignX (a component of Vector NTI Suite 6.0 from Informax, Inc.), which is an implementation of the ClustalW alignment algorithm [Thompson, J. D., D. G. Higgins, et al. (1994). Nucleic Acids Res 22(22): 4673-80.]. *See* Figure 2. The sections of the alignment for which the reference sequence (*E. cloaceae*) had a gap character were discarded, as only positions at which the reference sequence contained an amino acid were used.

The multiple sequence alignment of ampC was used to generate a profile using the method of Gribskov as described above except that a mutation probability matrix was used instead of the log-odds substitution matrix form used by Gribskov. The mutation probability matrix gives the probabilities that any given amino acid will mutate to each of the other amino acids in a given evolutionary interval. The mutation probability matrix PAM 32, which was generated from the PAM1 matrix as described [Dayhoff, M. et al. (1978) *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Washington), Vol. 5, Suppl. 3, pp. 345-358)], was used.

A distance-based constraint was applied to the scoring matrix to limit mutations to residues that are surface exposed and within 6 angstroms from the binding site of ligands in the *E. cloacae* ampC 3D structure. Specifically, the *E. cloacae* ampC crystal structure (Protein Database Base ID# 1BLS) and 6 *E. coli* ampC structures containing bound inhibitors or substrates (Protein Database Base structures 1C3B, 1FCM, 1FCN, 1FCO, 1FSW, 1FSY) were first loaded into the program MOE 2000.01 (Chemical Computing Group, Inc., Montreal Canada). Because each structure consists of a homodimer, one of the monomers and its associated ligand was deleted. Next, the main chains of all the structures containing bound ligands were aligned (0.4 angstroms RMS deviation) and all the water molecules were manually deleted. The main chains of all structures except the *E. cloacae* structure (1BLS) were then removed. The resulting structure consisted of the *E. cloacae* ampC molecule with all of the superimposed ligands from the other 6 ampC structures. All surface-exposed side chains (i.e, the beta carbon and additional atoms not in the backbone) in ampC with atoms within 6 angstroms of the ligand atoms were then selected for the IRL library. Five of the top